COLLABORATING TO AUTOMATE BIG DATA CLEANING: AN EXAMPLE USING LOCAL BIBLIOMETRIC DATA

Jana Carson, Senior Manager, Evaluation & Accountability, Institutional Analysis & Planning

Shannon Gordon, Bibliometrics & Research Impact Librarian, Library

April 6, 2017

#UWStaffConf2017

UNIVERSITY OF WATERLOO

2017 WATERLOO STAFF CONFERENCE
ORGANIZATIONAL & HUMAN DEVELOPMENT

AGENDA

- Background
- Our big data project
  - Collaborative power & impact
  - Looking ahead
- Other projects
- Discussion

BACKGROUND

BIBLIOMETRICS?

"A statistical or mathematical method for counting the number of academic publications, citations and authorship. *It is frequently used to measure academic output*. Data usually comes from an international database e.g. Thomson Reuters Web of Science or Elsevier-Scopus." (European Commission on Research and Innovation)

KEY COLLABORATIONS

UNIVERSITY OF WATERLOO

INITIATING COLLABORATIONS

Advisory Group

Working Group on Bibliometrics

White Paper Sub-Committee

EXAMPLE PROJECTS

- White Paper on Bibliometrics
- Discipline-level bibliometric analysis (BIG DATA!)
- Collaborative research partnerships
- Validating university rankings data
- Evidence for competitive applications

Consultation Fall 2015

Finalized Winter 2016

Added to UWSpace

Companion guide

RECOMMENDED PRACTICES FOR USERS OF BIBLIOMETRIC DATA

Analyze research outputs in the same way as conducting good research	Work from a basket of measures	Account for varying publication cultures
Involve those being evaluated in the process	Understand distinctions among bibliometric measures	Exercise caution when using journal impact rankings

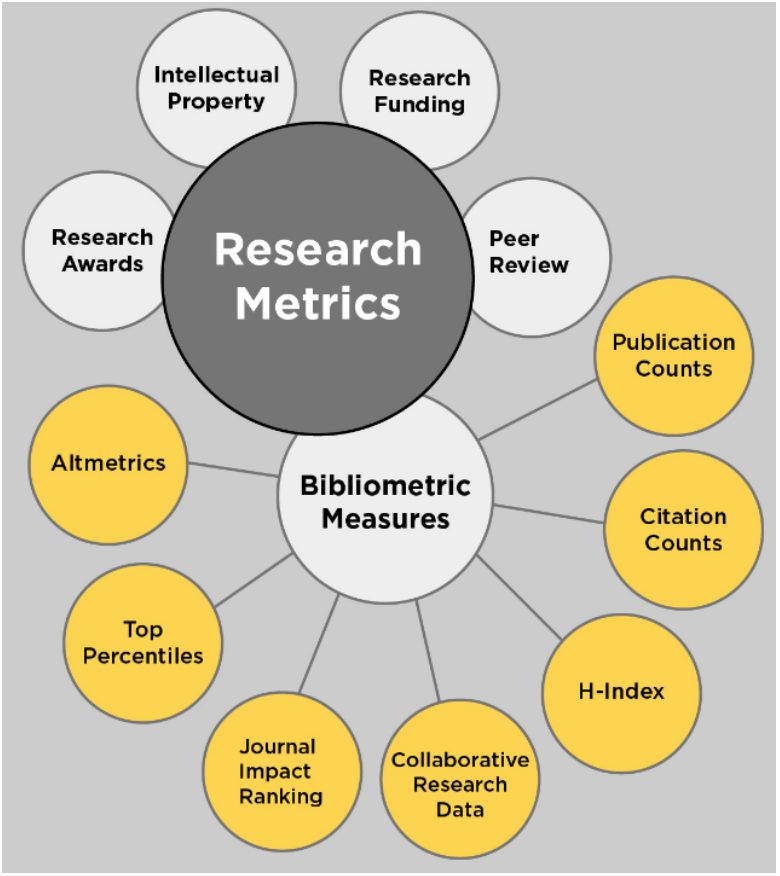
OUR BIG DATA PROJECT

COLLABORATING TO AUTOMATE BIG DATA CLEANING

COLLABORATORS

THE PROJECT

- Research area
  - Strategic highlights
  - Research funding
  - Research productivity



(University of Waterloo Working Group on Bibliometrics, 2016)

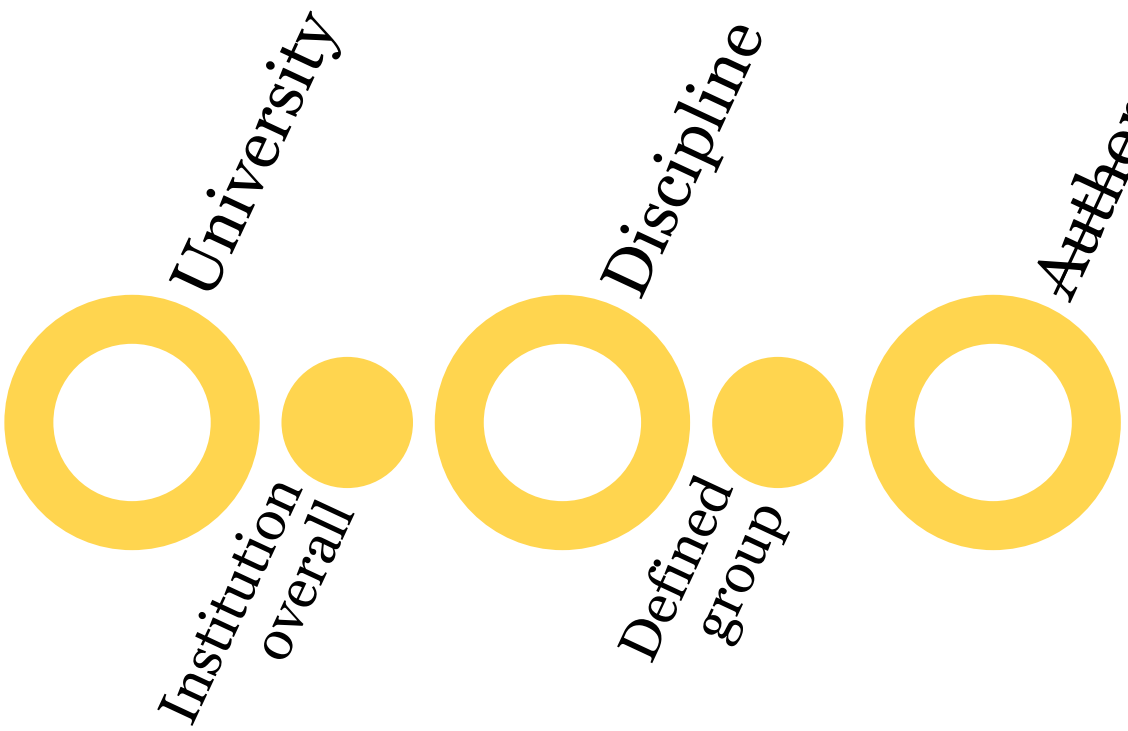
THE DATA

UNIVERSITY OF WATERLOO

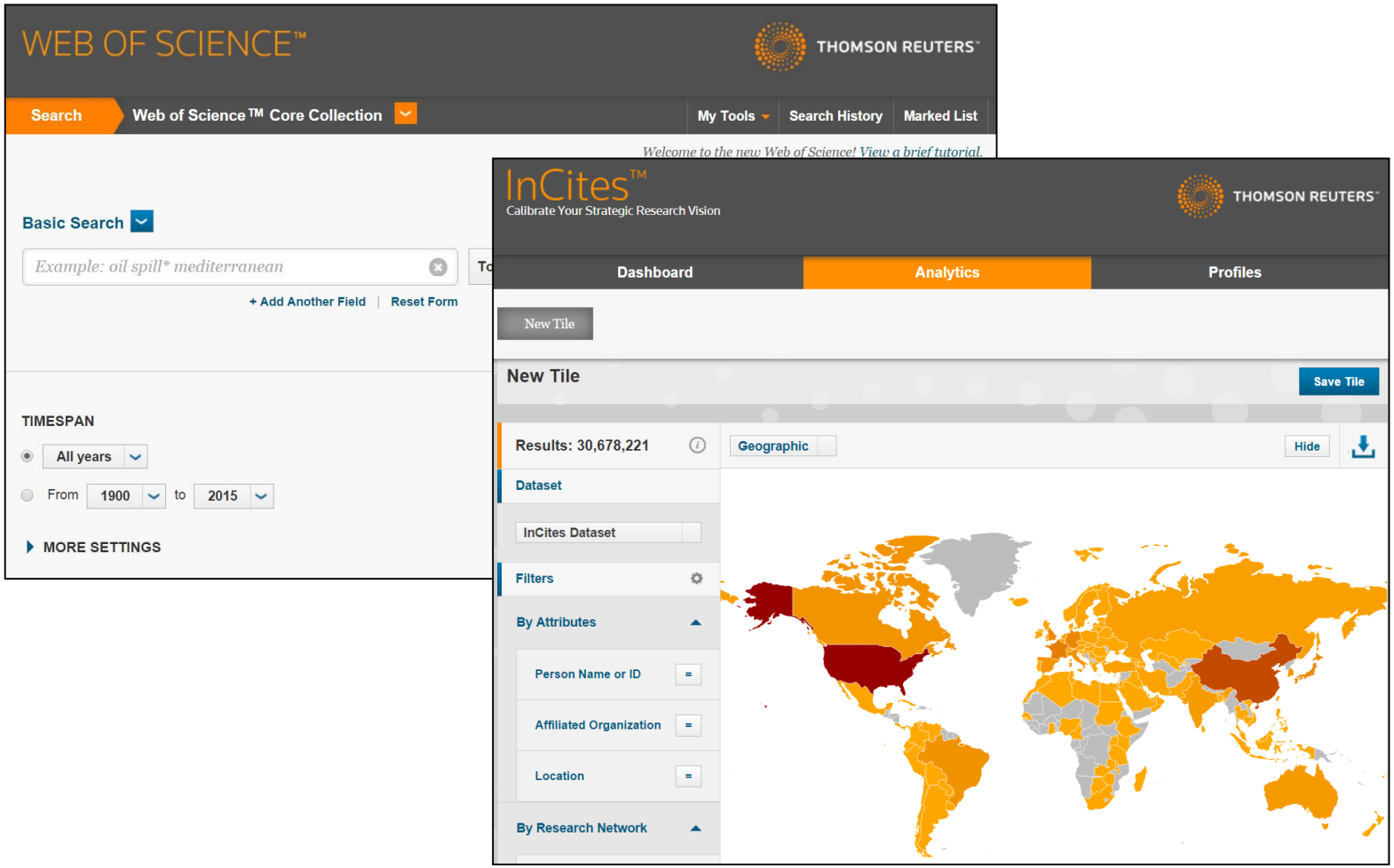
RESEARCH PRODUCTIVITY

Total publications (annual, 5-year, overall)	Total citations (annual, 5-year, overall)	Average # citations per publication
Median number of citations	Average # publications per member	Total publications with citations
Total publications with zero citations	Total publications with 100 or more citations	H-index

INTERNAL ANALYSIS



DATA SOURCE



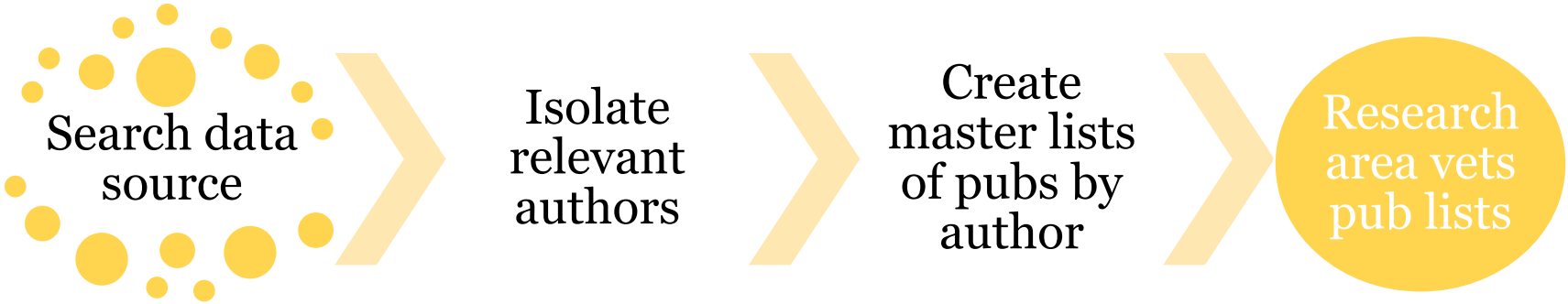
UNIVERSITY OF WATERLOO

OUR RAW DATA...EEK!

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Accession Number	DOI	Pubmed ID	Article Title	Link	Authors	Source	Volume	Issue	Pages	Publication Date	Times Cited	Journal Expected	Category Expected	Journal Normalized	Category Normalized	Percentile in Subject	Journal Impact
2	WOS:0003373502	10.1038/natu	MEDLINE::	Contextuality sup	https://ga	Howard, Mark; Wa	NATURE	510	7505	351-355	2014	27	38.58	3.29	0.7	8.21	1.37	41.46
3	WOS:0003440312	10.1103/Phy	MEDLINE::	Interferometric P	https://ga	Serbyn, M.; Knap,	PHYSICAL REVIEW	113	14		2014	27	9.85	3.29	2.74	8.21	1.37	7.51
4	WOS:0003323298	10.1103/PhysRevA.89.0		Improved entropi	https://ga	Coles, Patrick J.; P	PHYSICAL REVIEW	89	2		2014	18	3.35	2.9	5.37	6.21	0.96	2.81
5	WOS:0003305748	10.1088/1367-2630/16/		The resource theo	https://ga	Veitch, Victor; Mo	NEW JOURNAL OF	16			2014	18	4.19	3.29	4.3	5.47	3.06	3.56
6	WOS:0003331882	10.1103/PhysRevA.89.0		Casimir forces on	https://ga	Alhambra, Alvaro I	PHYSICAL REVIEW	89	3		2014	17	3.35	2.9	5.07	5.87	1.08	2.81
7	WOS:0003382844	10.1126/scie	MEDLINE::	Tunable fractiona	https://ga	Maher, Patrick; Wa	SCIENCE	345	6192	61-64	2014	17	30.97	3.29	0.55	5.17	3.52	33.61
8	WOS:0003351567	10.1007/s00220-014-19		Everything You Al	https://ga	Chitambar, Eric; Le	COMMUNICATION	328	1	303-326	2014	15	3.15	2	4.76	7.51	1.03	2.09
9	WOS:0003338000	10.1038/NPHOTON.201		Experimental thre	https://ga	Erven, C.; Meyer-S	NATURE PHOTONI	8	4	292-296	2014	12	31.57	2.89	0.38	4.16	2.7	32.39
10	WOS:0003319396	10.1103/Phy	MEDLINE::	Using Concatenat	https://ga	Jochym-O'Connor,	PHYSICAL REVIEW	112	1		2014	12	9.85	3.29	1.22	3.65	6.28	7.51
11	WOS:0003431452	10.1038/NPHOTON.201		Direct generation	https://ga	Hamel, Deny R.; Sh	NATURE PHOTONI	8	10	801-807	2014	11	31.57	2.89	0.35	3.81	3.39	32.39
12	WOS:0003359288	10.1103/Phy	MEDLINE::	Hamiltonian Lear	https://ga	Wiebe, Nathan; Gr	PHYSICAL REVIEW	112	19		2014	11	9.85	3.29	1.12	3.35	7.24	7.51
13	WOS:0003319525	10.1103/Phy	MEDLINE::	Laser Damage Hel	https://ga	Bugge, Audun Nys	PHYSICAL REVIEW	112	7		2014	10	9.85	3.29	1.02	3.04	8.38	7.51
14	WOS:0003345931	10.1103/PhysRevA.89.0		Quantum Hamiltc	https://ga	Wiebe, Nathan; Gr	PHYSICAL REVIEW	89	4		2014	10	3.35	2.9	2.99	3.45	4.15	2.81
15	WOS:0003372777	10.1103/Phy	MEDLINE::	Experimental Rea	https://ga	Li, Zhaokai; Zhou, I	PHYSICAL REVIEW	112	22		2014	9	9.85	3.29	0.91	2.74	9.7	7.51
16	WOS:0003463228	10.1021/nl5(	MEDLINE::	Polarization Entar	https://ga	Huber, Tobias; Pre	NANO LETTERS	14	12	7107-71	2014	9	14.57	4.62	0.62	1.95	9.12	13.59
17	WOS:0003444068	10.1103/PhysRevB.90.1		Quantum quench	https://ga	Serbyn, Maksym; F	PHYSICAL REVIEW	90	17		2014	9	4.35	4.5	2.07	2	12.98	3.74
18	WOS:0003440961	10.1088/1367-2630/16/		Randomized benc	https://ga	Wallman, Joel J.; F	NEW JOURNAL OF	16			2014	9	4.19	3.29	2.15	2.74	9.7	3.56
19	WOS:0003319468	10.1103/Phy	MEDLINE::	Topological Phase	https://ga	Papic, Z.; Abanin, I	PHYSICAL REVIEW	112	4		2014	9	9.85	3.29	0.91	2.74	9.7	7.51
20	WOS:0003412607	10.1103/PhysRevA.90.0		Nonclassical high	https://ga	Rundquist, Arman	PHYSICAL REVIEW	90	2		2014	8	3.35	2.9	2.39	2.76	6.89	2.81
21	WOS:0003321690	10.1103/Phy	MEDLINE::	Quantum Nonder	https://ga	Sathyamoorthy, Sa	PHYSICAL REVIEW	112	9		2014	8	9.85	3.29	0.81	2.43	11.27	7.51
22	WOS:0003323306	10.1103/PhysRevA.89.0		Quantum signalin	https://ga	Jonsson, Robert H.	PHYSICAL REVIEW	89	2		2014	8	3.35	2.9	2.39	2.76	6.89	2.81
23	WOS:0003321680	10.1103/PhysRevD.89.0		Echo of the quant	https://ga	Garay, Luis J.; Mart	PHYSICAL REVIEW	89	4		2014	7	6.22	5	1.12	1.4	22.11	4.64
24	WOS:0003446073	10.1088/0264-9381/31/		Entanglement in (	https://ga	Martin-Martinez, E	CLASSICAL AND QU	31	21		2014	7	3.68	4.26	1.9	1.64	13.39	3.17
25	WOS:0003388438	10.1088/1367-2630/16/		Experimental real	https://ga	Lu, Dawei; Brodut	NEW JOURNAL OF	16			2014	7	4.19	3.29	1.67	2.13	13.39	3.56
26	WOS:0003345133	10.1088/1367-2630/16/		Memory-assisted	https://ga	Panayi, Christiana;	NEW JOURNAL OF	16			2014	7	4.19	3.29	1.67	2.13	13.39	3.56
27	WOS:0003394461	10.1103/PhysRevB.90.0		Single-mode appr	https://ga	Repellin, Cecile; N	PHYSICAL REVIEW	90	4		2014	7	4.35	4.5	1.61	1.56	18.04	3.74
28	WOS:0003412885	10.1103/PhysRevD.90.0		Spacetime effects	https://ga	Bruschi, David Edw	PHYSICAL REVIEW	90	4		2014	7	6.22	5	1.12	1.4	22.11	4.64
29	WOS:0003386343	10.1103/Phy	MEDLINE::	Ultrafast and Faul	https://ga	Muralidharan, Sree	PHYSICAL REVIEW	112	25		2014	7	9.85	3.29	0.71	2.13	13.39	7.51
30	WOS:0003343004	10.1038/nco	MEDLINE::	An experimental	https://ga	Erven, C.; Ng, N.; G	NATURE COMMUN	5			2014	6	10.31	2.47	0.58	2.43	12.03	11.47
31	WOS:0003476137	10.1038/nco	MEDLINE::	Equivalence of wa	https://ga	Coles, Patrick J.; Ka	NATURE COMMUN	5			2014	6	10.31	3.29	0.58	1.82	15.89	11.47
32	WOS:0003369153	10.1103/PhysRevA.89.0		Quantum fingerpr	https://ga	Arrazola, Juan Mig	PHYSICAL REVIEW	89	6		2014	6	3.35	2.9	1.79	2.07	12.03	2.81
33	WOS:0003331072	10.1103/PhysRevD.89.0		Scalar Green func	https://ga	Yang, Huan; Zhang	PHYSICAL REVIEW	89	6		2014	6	6.22	5	0.96	1.2	26.41	4.64
34	WOS:0003412370	10.1103/PhysRevB.90.0		Solvable models f	https://ga	Papic, Z.	PHYSICAL REVIEW	90	7		2014	6	4.35	4.5	1.38	1.33	21.63	3.74
35	WOS:0003322216	10.1103/PhysRevA.89.0		Tractable simulati	https://ga	Puzzuoli, Daniel; G	PHYSICAL REVIEW	89	2		2014	6	3.35	2.9	1.79	2.07	12.03	2.81

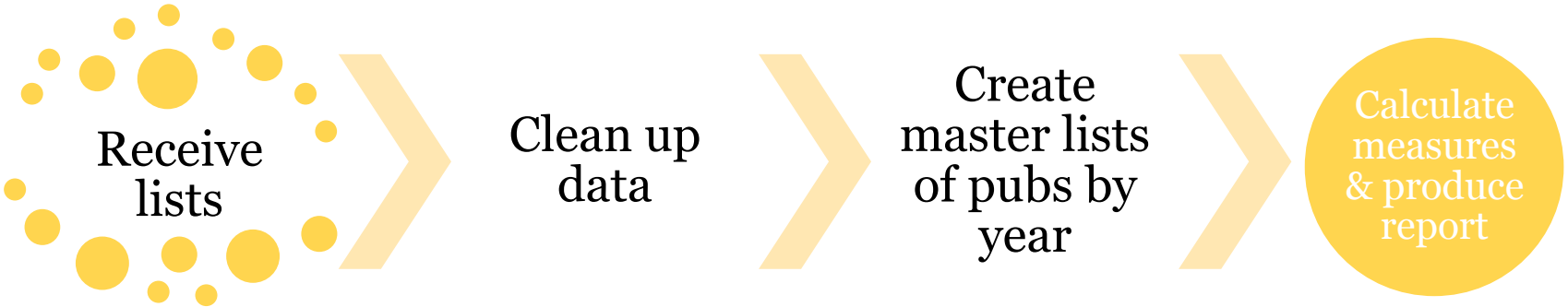
COLLABORATIVE DATA STAGES

INPUT DATA WORKFLOW



Author A
Publication 1
Publication 2
Publication 3

OUTPUT DATA WORKFLOW



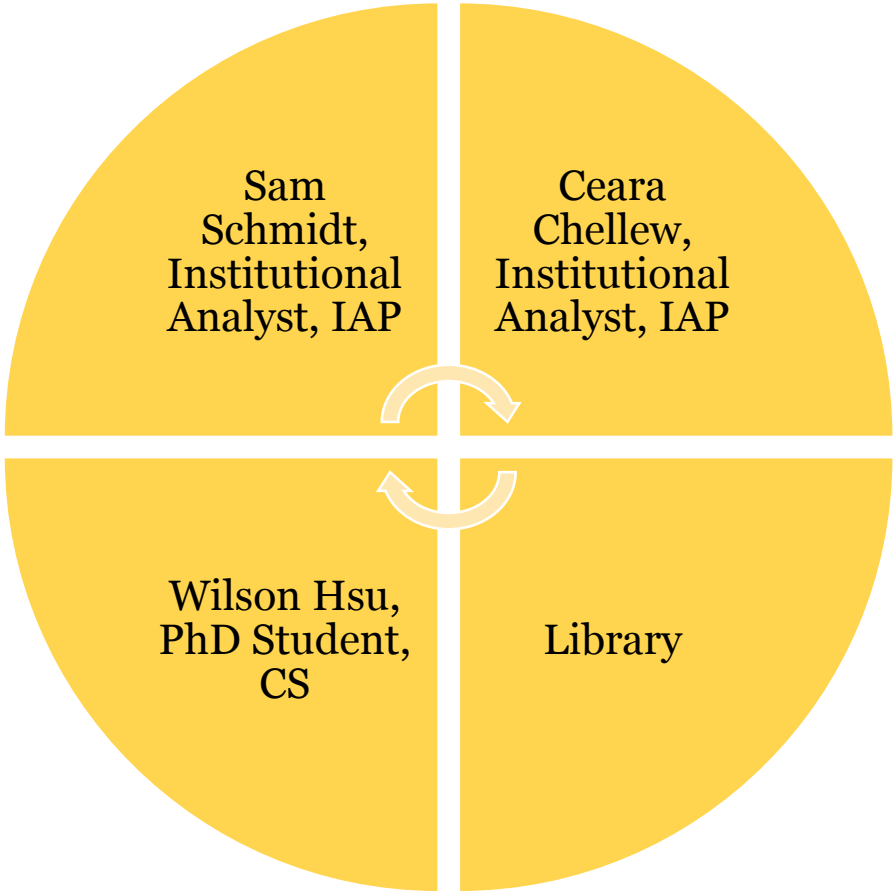
Year A
Publication 1
Publication 2
Publication 3

THE CHALLENGES

- Isolating defined data points
  - Author names
  - Publication worksheets
- Updating needed data points
- Calculating various data points
- Cleaning up all data & generating final datasets

COLLABORATIVE POWER

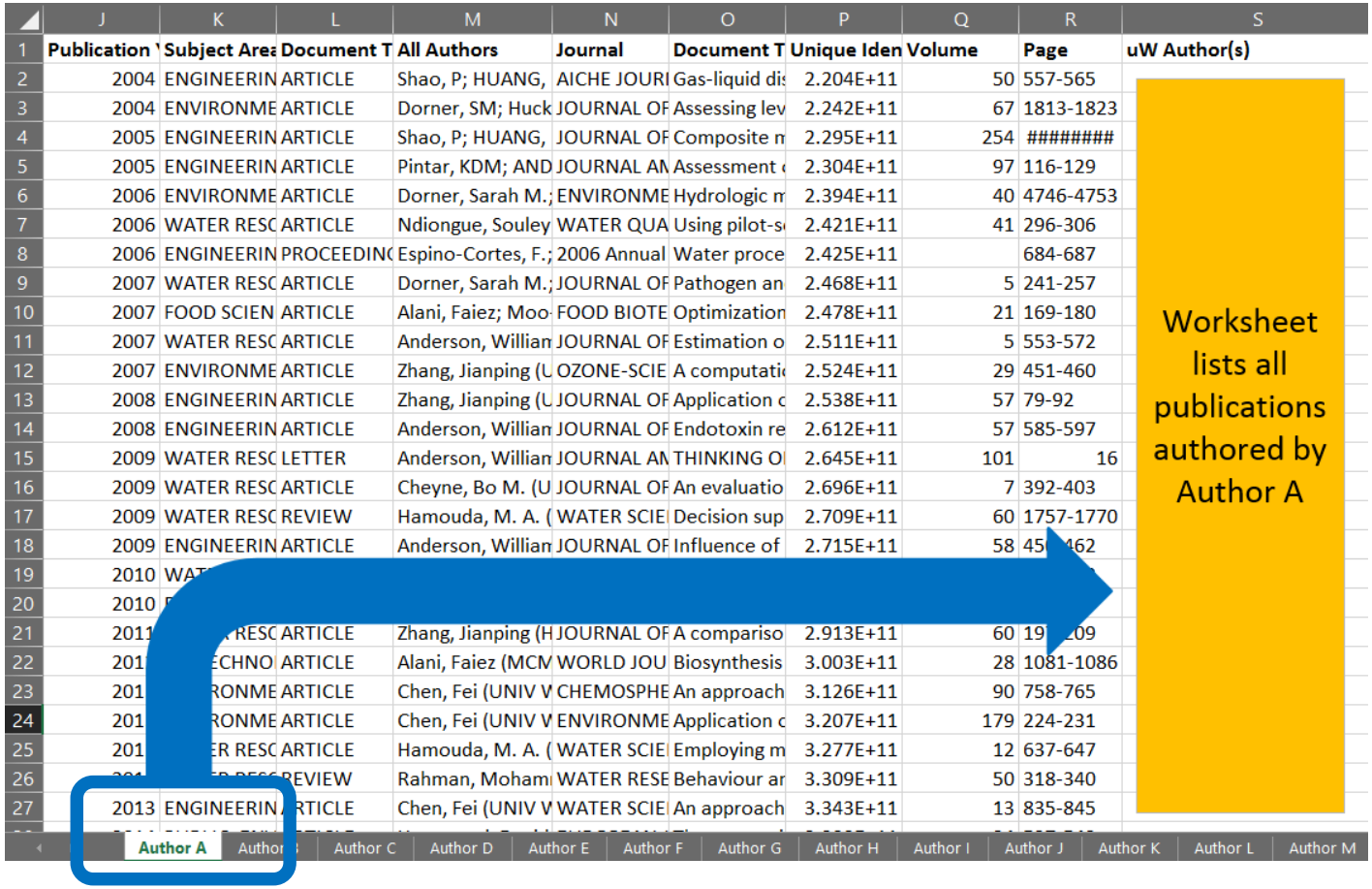
TEAMING UP TO FIND A SOLUTION



COLLABORATIVE SOLUTION #1

Challenge	Manual process	Automated solution	Time savings
Isolating defined data points in a dataset **~600 names** **~600 lists**	Names: ~20 hr/area Pub lists: ~6 hr/area	Python script isolates relevant author names Python script generates author-level lists	~130 manual hours

ISOLATING DEFINED DATA POINTS



	J	K	L	M	N	O	P	Q	R	S
1	Publication '	Subject Area	Document T	All Authors	Journal	Document T	Unique Iden	Volume	Page	uW Author(s)
2	2004	ENGINEERIN	ARTICLE	Shao, P; HUANG,	AICHE JOURI	Gas-liquid dis	2.204E+11	50	557-565	
3	2004	ENVIRONME	ARTICLE	Dorner, SM; Huck	JOURNAL OF	Assessing lev	2.242E+11	67	1813-1823	
4	2005	ENGINEERIN	ARTICLE	Shao, P; HUANG,	JOURNAL OF	Composite m	2.295E+11	254	########	
5	2005	ENGINEERIN	ARTICLE	Pintar, KDM; AND	JOURNAL AM	Assessment	2.304E+11	97	116-129	
6	2006	ENVIRONME	ARTICLE	Dorner, Sarah M.;	ENVIRONME	Hydrologic m	2.394E+11	40	4746-4753	
7	2006	WATER RESC	ARTICLE	Ndiongue, Souley	WATER QUA	Using pilot-s	2.421E+11	41	296-306	
8	2006	ENGINEERIN	PROCEEDING	Espino-Cortes, F.;	2006 Annual	Water proce	2.425E+11		684-687	
9	2007	WATER RESC	ARTICLE	Dorner, Sarah M.;	JOURNAL OF	Pathogen an	2.468E+11	5	241-257	
10	2007	FOOD SCIEN	ARTICLE	Alani, Faiez; Moo	FOOD BIOTE	Optimization	2.478E+11	21	169-180	
11	2007	WATER RESC	ARTICLE	Anderson, Willian	JOURNAL OF	Estimation o	2.511E+11	5	553-572	
12	2007	ENVIRONME	ARTICLE	Zhang, Jianping (U	OZONE-SCIE	A computati	2.524E+11	29	451-460	
13	2008	ENGINEERIN	ARTICLE	Zhang, Jianping (U	JOURNAL OF	Application c	2.538E+11	57	79-92	
14	2008	ENGINEERIN	ARTICLE	Anderson, Willian	JOURNAL OF	Endotoxin re	2.612E+11	57	585-597	
15	2009	WATER RESC	LETTER	Anderson, Willian	JOURNAL AM	THINKING OI	2.645E+11	101	16	
16	2009	WATER RESC	ARTICLE	Cheyne, Bo M. (U	JOURNAL OF	An evaluatio	2.696E+11	7	392-403	
17	2009	WATER RESC	REVIEW	Hamouda, M. A. (	WATER SCIE	Decision sup	2.709E+11	60	1757-1770	
18	2009	ENGINEERIN	ARTICLE	Anderson, Willian	JOURNAL OF	Influence of	2.715E+11	58	45 462	
19	2010	WAT								
20	2010									
21	2011	RESC	ARTICLE	Zhang, Jianping (H	JOURNAL OF	A compariso	2.913E+11	60	19 209	
22	201	ECHNO	ARTICLE	Alani, Faiez (MCM	WORLD JOU	Biosynthesis	3.003E+11	28	1081-1086	
23	201	RONME	ARTICLE	Chen, Fei (UNIV V	CHEMOSPHE	An approach	3.126E+11	90	758-765	
24	201	RONME	ARTICLE	Chen, Fei (UNIV V	ENVIRONME	Application c	3.207E+11	179	224-231	
25	201	ER RESC	ARTICLE	Hamouda, M. A. (	WATER SCIE	Employing m	3.277E+11	12	637-647	
26			REVIEW	Rahman, Moham	WATER RESE	Behaviour ar	3.309E+11	50	318-340	
27	2013	ENGINEERIN	RTICLE	Chen, Fei (UNIV V	WATER SCIE	An approach	3.343E+11	13	835-845	

Worksheet lists all publications authored by Author A

Author A | Author B | Author C | Author D | Author E | Author F | Author G | Author H | Author I | Author J | Author K | Author L | Author M

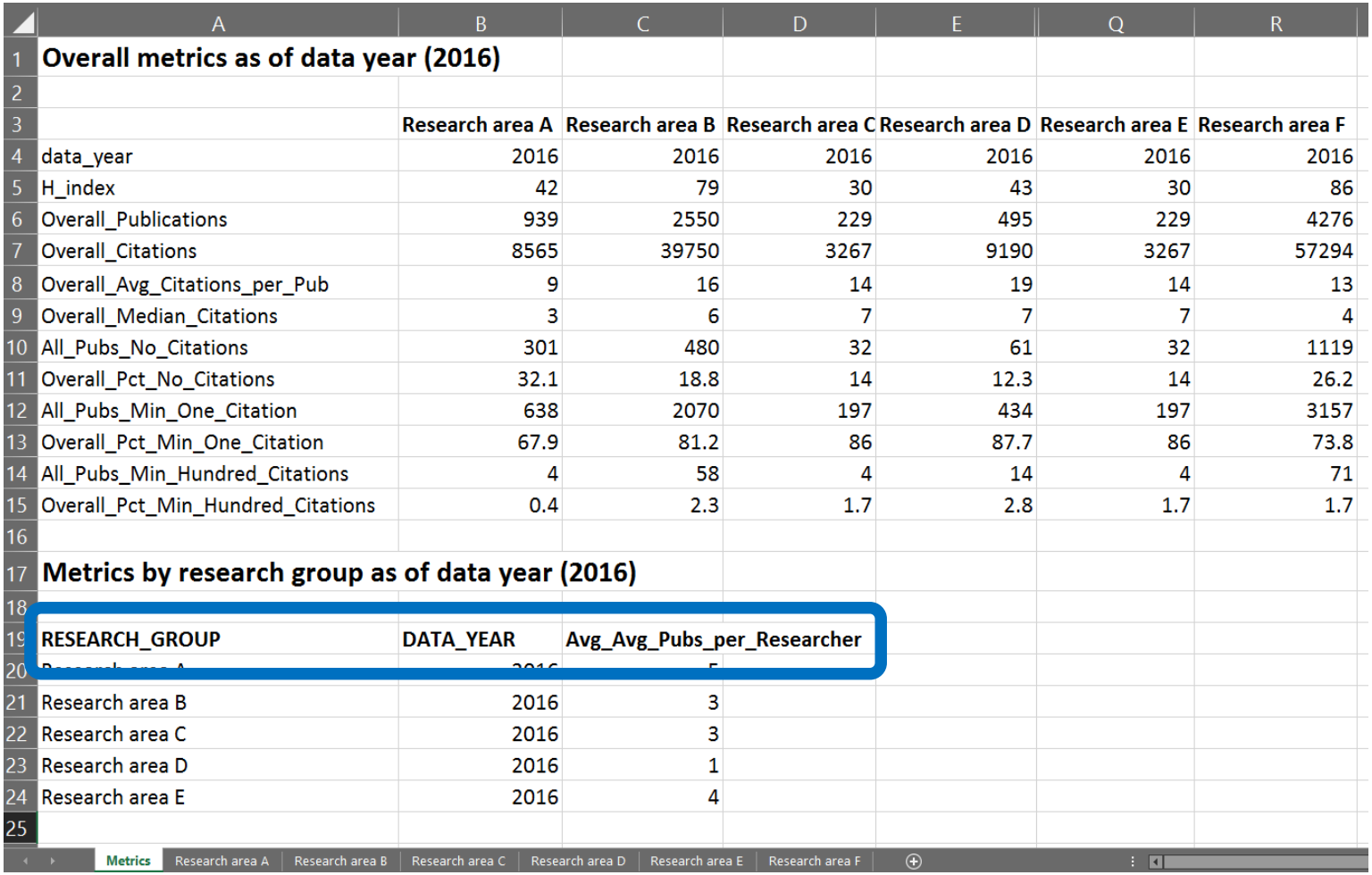
COLLABORATIVE SOLUTION #2

Challenge	Automated solution
Bulk editing data points in several datasets **~25,000 pubs** **Updating 'old' data & adding 'new' data**	Library gives IAP: - Current metrics for 'old' pub data - Python output of 'new' pub data IAP then runs an ETL process to: - Update 'old' metrics and add 'new' metrics - Data integrity checks - Produce pub reports by year

COLLABORATIVE SOLUTION #3

Challenge	Manual process	Automated solution	Time savings
Calculating final data points **~200 bibliometric measures**	Calculated via Excel formulas	SQL server database calculates ~200 data points for all research areas	~50 hours

AUTOMATED REPORTS

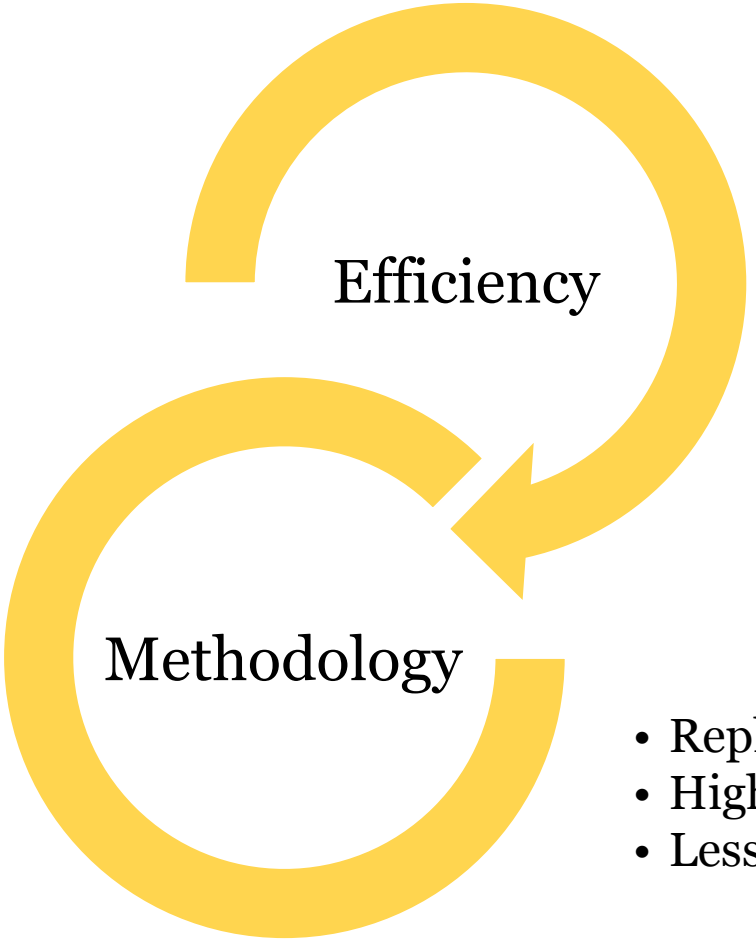


	A	B	C	D	E	Q	R
1	**Overall metrics as of data year (2016)**						
2							
3		**Research area A**	**Research area B**	**Research area C**	**Research area D**	**Research area E**	**Research area F**
4	data_year	2016	2016	2016	2016	2016	2016
5	H_index	42	79	30	43	30	86
6	Overall_Publications	939	2550	229	495	229	4276
7	Overall_Citations	8565	39750	3267	9190	3267	57294
8	Overall_Avg_Citations_per_Pub	9	16	14	19	14	13
9	Overall_Median_Citations	3	6	7	7	7	4
10	All_Pubs_No_Citations	301	480	32	61	32	1119
11	Overall_Pct_No_Citations	32.1	18.8	14	12.3	14	26.2
12	All_Pubs_Min_One_Citation	638	2070	197	434	197	3157
13	Overall_Pct_Min_One_Citation	67.9	81.2	86	87.7	86	73.8
14	All_Pubs_Min_Hundred_Citations	4	58	4	14	4	71
15	Overall_Pct_Min_Hundred_Citations	0.4	2.3	1.7	2.8	1.7	1.7
16							
17	**Metrics by research group as of data year (2016)**						
18							
19	**RESEARCH_GROUP**	**DATA_YEAR**	**Avg_Avg_Pubs_per_Researcher**				
20	[illegible]	[illegible]	[illegible]				
21	Research area B	2016	3				
22	Research area C	2016	3				
23	Research area D	2016	1				
24	Research area E	2016	4				
25							

Metrics | Research area A | Research area B | Research area C | Research area D | Research area E | Research area F

COLLABORATIVE IMPACT

COLLABORATIVE IMPACT



- Automation saved ~200 hours of manual processes
- **Key players**
  - **IAP Analysts**
  - **CS Grad Student**
  - **Library**

- Replicable
- High quality, robust data
- Less prone to error

VALUE OF DOCUMENTATION

- Significant spot checking
- Data integrity analysis
- Process documentation
  - Searching methodology
  - Technical processes
- ~30 data dictionaries

COMMON UNDERSTANDING

- ~30 data dictionaries
- Reliable
- Replicable
- Facilitates systematic understanding

Data dictionary criteria	
Project Title	
Metric Title	
IAP Technical Name	
Technical Definition	
Description	
Purpose	
Base Time	
Category	
Details	Inclusions
	Exclusions
	Time Frame
	Date Available
	Source
Owner	Performance Owner
	Primary Data Owner
	Primary Data Owner Contact
	Secondary Data Owner
Reported In	
Notes	

LOOKING AHEAD

NEXT STEPS

- Continued collaboration
- Ongoing tweaking of Python script
- Web of Science & InCites APIs
- Online database to enhance vetting experience
- Bibliometric data warehouse
- Application to other discipline-level analyses

BARRIERS

Capacity to support automation needs
Identifying campus expertise

OTHER PROJECTS

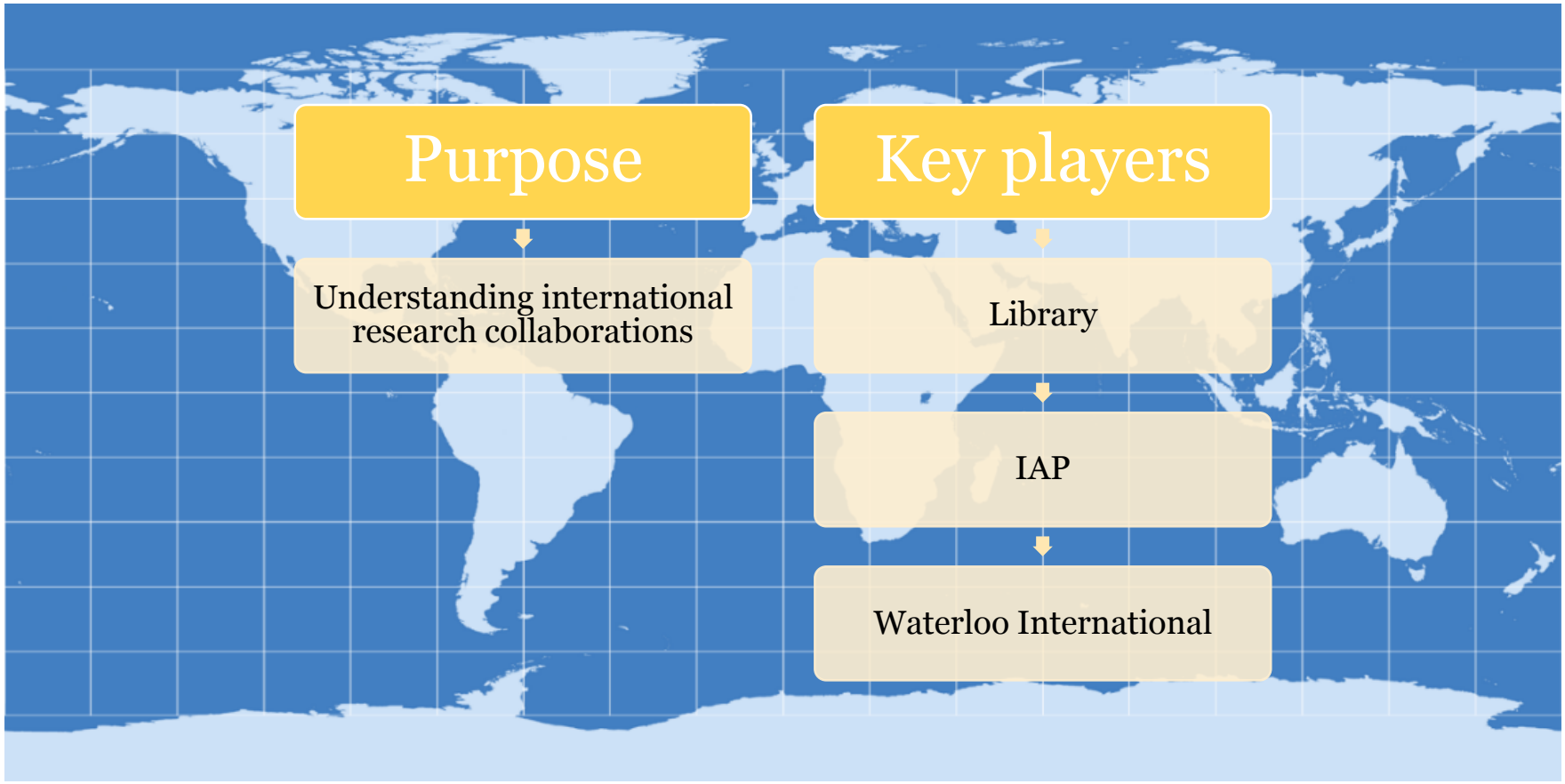
LEVERAGING COLLABORATIONS FOR OTHER PROJECTS

Informing university partnership profiles

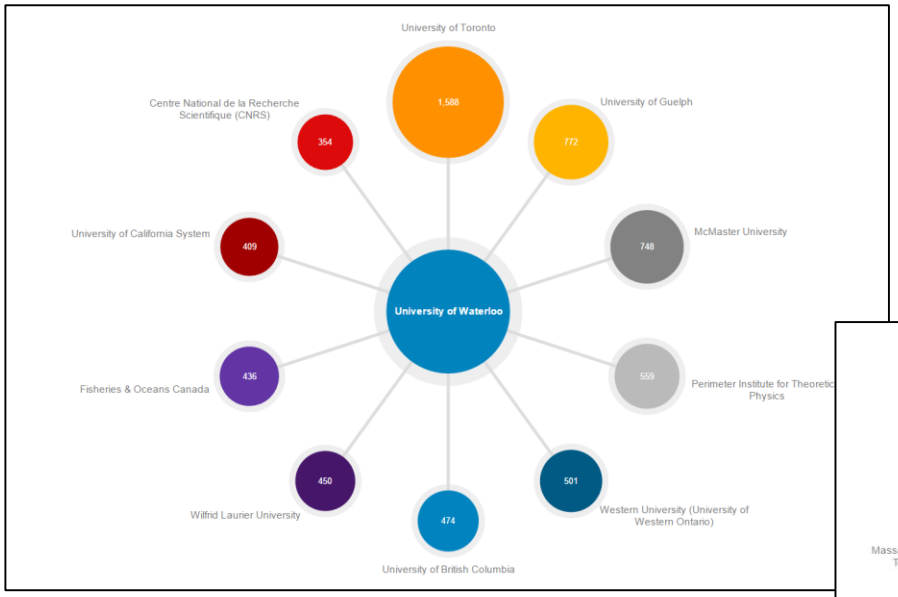
Validating university rankings data

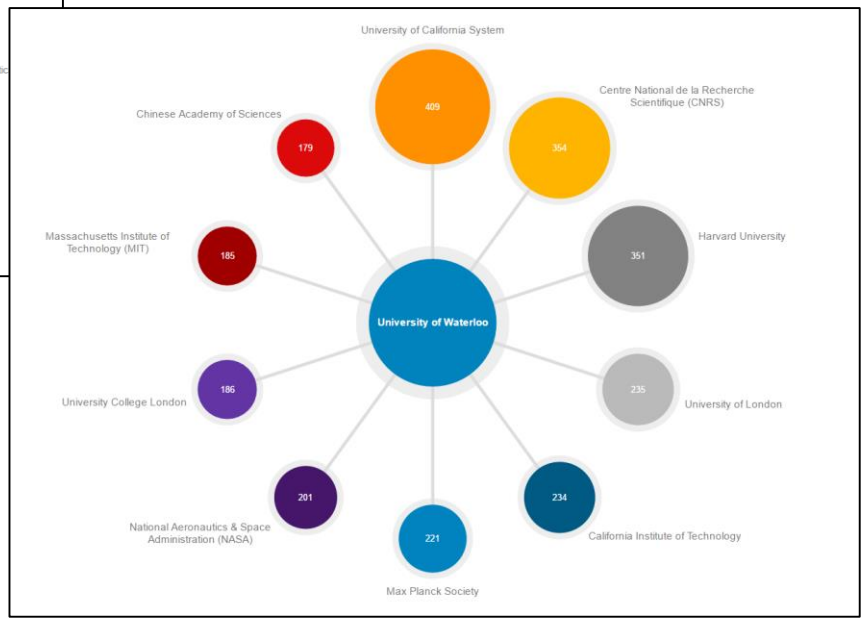
Evidence for competitive grant applications

COLLABORATIVE RESEARCH



WATERLOO'S RESEARCH COLLABORATIONS





UNIVERSITY OF WATERLOO

VALIDATING UNIVERSITY RANKINGS DATA

Purpose

Deeper understanding of rankings data

Key players

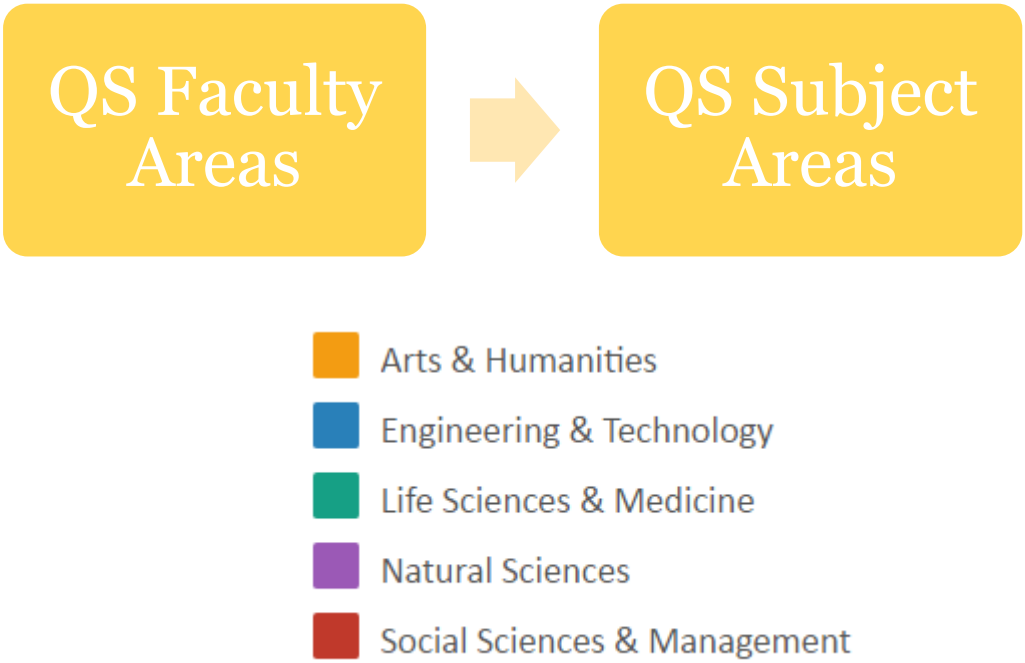
Faculty contacts

IAP

Library

EXPLORING DATA TRENDS

- How does subject-level data align with Departments and/or Faculties?



REPLICATING DATA TRENDS

- Can Waterloo data be replicated?

Publication and Citation Totals for University of Waterloo Authored Items Indexed in Scopus, for Items Published 2010-2014									
Year	Total Publications 2010-2014			Total Citations 2010-2014			Total Citations 2010-2014, Self-Citations Excluded		
	QS Data	Replicated Data	Difference	QS Data	Replicated Data	Difference	QS Data	Replicated Data	Difference
2010	3,281	3,248	-33	Unknown (Fact File does not provide citation breakdown)	1,521	NA	Unknown (Fact File does not provide citation breakdown)	Unknown (Currently unable to replicate due to Scopus issue)	NA
2011	3,475	3,384	-91		7,523	NA			NA
2012	3,631	3,593	-38		17,233	NA			NA
2013	3,708	3,676	-32		28,203	NA			NA
2014	3,463	3,674	211		39,325	NA			NA
Totals	**17,558**	**17,575**	**17**	**85,719**	**93,805**	**8,086**	**65,681**		**NA**
Notes	The difference in total publications is only 17 items. This could be			The difference of ~8,000 citations may be attributed to changes in the 2015			Given the size of the citation data request, Scopus only provides an		

THANK YOU!

jana.carson@uwaterloo.ca

shannon.gordon@uwaterloo.ca

WORKS CITED

- University of Waterloo Working Group on Bibliometrics. (Winter 2016). *White paper: Measuring research output through bibliometrics*. Retrieved from https://uwspace.uwaterloo.ca/bitstream/handle/10012/10323/Bibliometrics%20White%20Paper%202016%20Final_March2016.pdf?sequence=4&isAllowed=y